### 包装系统

#### 数据集

Database由Yiding提供，可以找她要

#### 数据清洗

目前数据的清洗逻辑还在梳理中，现有逻辑有：

1. 删除LENGTH(MM) WIDTH(MM) HEIGHT(MM) 或 UL没有的数据

2.删除 df = df.loc[-((df["EXP"] == "Y") | (df["S"] == "N"))]

3.删除长宽高为1的数据；

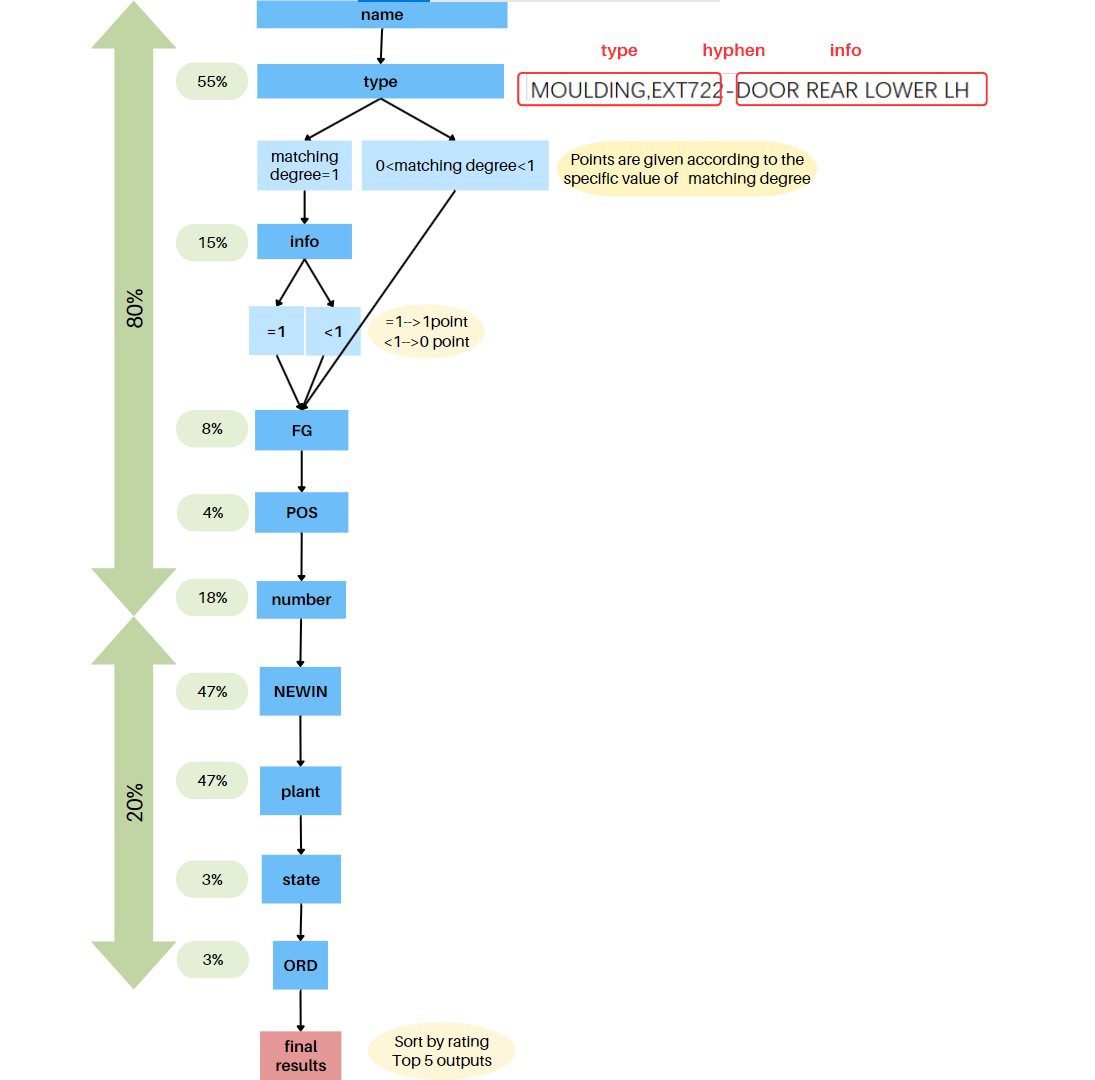
4.找到包装箱型（LC）相同但长宽高不同的数据，交由业务选择正确数据。（进行中）

5.

#### 模型建立

##### 流程图

最新逻辑



该方法目前业务方可以自行调参

##### Machine Learning

1. 特征工程

首先要想应用机器学习，得先量化特征。原本database里面的特征都是类别变量，无法直接应用到sklearn中的分类模型中。再加上类别变量的类别都很多，上百上千，因此也不适合用one-hot-encoding和labelencoding，所以我选择了target-encoding。但是target-encoding是基于每个类别目标变量的均值赋值的，均值不具有代表性，可能还会有偏差，于是进一步优化，使用beta-target-encoding（[Avito Demand Prediction Challenge | Kaggle](https://www.kaggle.com/c/avito-demand-prediction/discussion/60059)|[Beta Target Encoding | Kaggle](https://www.kaggle.com/code/mmotoki/beta-target-encoding)）对特征进行encoding。

对于要分类的类别变量，single\_volume，采用labelencoding的方式进行编码

2.训练集测试集划分

首先我尝试了train\_test\_split()，但考虑到类别不均衡问题，尝试使用分层划分，划分训练集和测试集，即从每一类中选取0.1作为训练集，保证训练集中包含所有的类别。然后针对类别不均衡问题，采用随机过采样法(over sampling)对训练集进行随机过采样。各个处理方法结果如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 分层划分 | 随机过采样 | 训练集 1vs1 完全一致(%) | 测试集 1v1 <0.3(%) |
| √ | √ | 0.961 | 0.61 |
| √ | × | 0.959 | 0.63 |
| × | × | 0.957 | 0.66 |
| × | √ | 0.958 | 0.63 |

结论是，不如不处理，越处理过拟合越严重。使用train\_test\_split()即可。。。。。。

1. 模型训练：

我目前只采用了RandomForest(n\_estimators = 15)，后续可以尝试LightGBM等模型。

1. 改进方向

从结果来看，这个模型目前最大的问题就是过拟合，如何解决过拟合问题是未来探索的方向。

还有一个问题是，随机森林在训练过程中会根据训练集自动为每个特征赋权，但根据业务知识，以及错误的结果分析来看，NAME和PLANT的指标重要性需要被加强，如何加强？